Amino Acids (2008) 35:607–614
DOI 10.1007/s00726-008-0089-7

ORIGINAL ARTICLE

Use of tetrapeptide signals for protein secondary-structure prediction

Yonge Feng · Liaofu Luo

Received: 15 December 2007 / Accepted: 4 March 2008 / Published online: 23 April 2008

Abstract This paper develops a novel sequence-based method, tetra-peptide-based increment of diversity with quadratic discriminant analysis (TPIDQD for short), for protein secondary-structure prediction. The proposed TPIDQD method is based on tetra-peptide signals and is used to predict the structure of the central residue of a sequence fragment. The three-state overall per-residue accuracy (Q_3) is about 80% in the threefold cross-validated test for 21-residue fragments in the CB513 dataset. The accuracy can be further improved by taking long-range sequence information (fragments of more than 21 residues) into account in prediction. The results show the tetra-peptide signals can indeed reflect some relationship between an amino acid's sequence and its secondary structure, indicating the importance of tetra-peptide signals as the protein folding code in the protein structure prediction.

Keywords Protein secondary-structure prediction · Tetra-peptide structural words · Increment of diversity · Quadratic discriminant analysis · Boundary correction · Long-range interaction

Electronic supplementary material The online version of this article (doi:10.1007/s00726-008-0089-7) contains supplementary material, which is available to authorized users.

Y. Feng · L. Luo (✉)
Laboratory of Theoretical Biophysics,
Faculty of Science and Technology,
Inner Mongolia University, Hohhot 010021, China
e-mail: lolfcm@mail.imu.edu.cn

Y. Feng
e-mail: fengyonge@163.com

Introduction

The prediction of protein structure and function from amino acid sequences is one of the most important problems in molecular biology. The field of protein structure prediction began even before the first protein structures were actually solved (Pauling et al. 1951). The secondary structure is the basis for the spatial structure of a protein. In terms of structure formation, the secondary structure is formed at the early stage of protein folding. Therefore, it is reasonable to study the protein secondary structure as the first and the most important step of three-dimensional structure prediction. Previous methods for secondary-structure prediction were based on single-residue statistics (Chou and Fasman 1974; Garnier et al. 1978) and provided generally poor accuracy. Then, a significant improvement was made with the PHD (Rost and Sander 1993) method, which used evolutionary information from multiple sequence alignments in the three-level neural networks.

The current prediction methods commonly employ machine learning techniques, such as neural networks (Jones 1999; Petersen et al. 2000; Dor and Zhou 2007), hidden Markov models (Karplus et al. 1998; Lin et al. 2005) and support vector machines (Ward et al. 2003; Guo et al. 2004), and have achieved an accuracy of Q_3 between 75 and 80%. Moreover, the accuracy can be further improved if the structure-based sequence alignments between high-homologue proteins are included as part of the prediction process (Montgomerie et al. 2006). On the other hand, along with secondary-structure prediction, several methods have been developed to predict protein structure class, protein-protein interaction and signal peptide (Chou 1995; Chou and Maggiora 1998; Chou and Zhang 1994; Cai et al. 2000; Chou and Cai

2006; Xiao et al. 2006; Zhang et al. 2007; Ding et al. 2007; Chou and Shen 2007c; Shen and Chou 2007b). The methodology development in predicting protein structural classification (Chou 2005a) has also greatly stimulated the areas for predicting the other attributes of proteins (Cedano et al. 1997; Chen et al. 2007; Chou 2001, 2005b; Chou and Elrod 1999; Chou and Shen 2007a, b; Guo et al. 2006b; Liu et al. 2005; Shen and Chou 2007a; Wang et al. 2004, 2005; Wen et al. 2006; Zhang et al. 2006a).

The empirical prediction of protein secondary structure essentially consists of two approaches: one is the direct sequence and structure comparison between high-homologue proteins by mapping the structure of known homologues onto the query protein's sequence; another is based on the search and exploitation of the information of sequence-structure pattern and on the development of an algorithm for structure classification. The latter approach is more important for the understanding of the law governing the sequence-structure relations in proteins and for the prediction of secondary structure of low-homologue proteins. In this article we concentrate on the latter. We introduce a novel method, tetra-peptide-based increment of diversity with quadratic discriminant analysis (TPIDQD for short), for protein secondary-structure prediction. TPIDQD is based on tetra-peptide signals (called tetra-peptide structural words). Tetra-peptide is the structural unit of alpha helix in the sense that the hydrogen bonds in the helix connect the J-th and $(J + 4)$-th residues, and they play a crucial role in the formation of the regular structure. It has been estimated that 60–70% of tetra-peptides encode the specific structure (Rackovsky 1993). These tetra-peptides can be regarded as the protein folding code. To test the method and facilitate comparison with previous studies, TPIDQD is tested on the CB513 (Cuff and Barton 1999) dataset and yields a higher accuracy. The success of the TPIDQD algorithm demonstrates the importance of tetra-peptide structural words in protein structure prediction.

Materials and methods

Dataset and definition of protein secondary structure

We have selected the non-homologous CB513 (Cuff and Barton 1999) dataset with sequence identity of less than 25% for secondary-structure prediction. The automatic assignments of secondary structure to experimentally determined 3D structure are usually performed using DSSP (Kabsch and Sander 1983), STRIDE (Frishman and Argos 1995), and DEFINE (Richards and Kundrot 1988). The DSSP assignments divide the secondary structure into eight categories: H (α-helix), E (extended β-strand), G (3_{10} helix), I (π-helix), B (isolated β-strand), T (turns), S (bend) and "_"(coil). Reducing eight classes to the three classes of helix (H), sheet (E), and coil (C) is an important step in the encoding of structure data (Rost and Sander 2000). Generally, two popular reduction methods are (1) H, G, and I to H; E to E; all others to C (Rost and Sander 1993), and (2) H, G to H; B, E to E; all others to C (Hua and Sun 2001). In CB513, method (2) yields many discrete states such as CEC, CEH, and HEC, that method (1) does not. So, we adopt the reduction method (1) in this study.

Definition of tetra-peptide structural words

Sliding a window of four residues along all protein sequences, one obtains the total occurrence frequency of a given tetra-peptide in the CB513 dataset, which will be denoted by N. In our statistics, only tetra-peptides with $N \geq 2$ are considered. Although a particular tetra-peptide with $N \geq 2$ may not occur in a given protein, we found each protein in CB513 dataset contains enough tetra-peptide signals with $N \geq 2$. Suppose the tetra-peptide occurs in structure j for n_j times, where $j = \alpha\alpha\alpha\alpha$, $\beta\beta\beta\beta$, $cccc$, $\alpha\alpha cc$, $\beta\beta cc$, $cc\alpha\alpha$, $cc\beta\beta$, $\alpha\alpha\alpha c$, $\beta\beta\beta c$, αccc, βccc, $ccc\alpha$, $ccc\beta$, $c\alpha\alpha\alpha$, $c\beta\beta\beta$. If its occurrence in structure j is a stochastic event, then the probability of the tetra-peptide occurring in this structure n_j or more times will be

$$1 - CL_j = \sum_{n \geq n_j} \frac{N!}{n!(N-n)!} p_j^n (1 - p_j)^{N-n} \quad (1)$$

where the sum in Eq. 1 is taken from n_j to N. In Eq. 1, p_j is defined by the relative frequency of structure j in the database, namely, $p_j = m_j/M$, here M is the total occurrence frequency of all tetra-peptides in the dataset and m_j their occurrence frequency in structure j. The details of parameters M, m_j, and p_j in the CB513 database are listed in Appendix 1. As Eq. 1 is a small quantity, the tetra-peptide occurring in structure j for n_j times should not be random. The confidence level of this statement is CL_j. For example, if the frequency of a tetra-peptide occurring in the j-th structure n_j satisfies Eq. 1 with $CL_j \geq 95\%$, then we say that the tetra-peptide is a j-type structural word with 95% confidence level.

Diversity and increment of diversity (ID)

Let n_i be the absolute frequency of the i-th category of some feature variable; there are t categories corresponding to a t-dimensional space (called category space). Set $X : \{n_i | i = 1, \cdots, t\}$ as the source of diversity. The measure of diversity (Laxton 1978; Li and Lu 2001) as a function of source X is defined by

$$D(X) = N \log\ N - \sum_{i=1}^{t} n_i \log\ n_i \quad (2)$$

$$\left(N = \sum_{i=1}^{t} n_i\right).$$

In general, for two sources of diversity in the same space of t dimensions, X:{n_1,n_2,...,n_t} and Y:{m_1,m_2,...,m_t}, the increment of diversity (Xu 1999) is defined by

$$ID(X,Y) = D(X+Y) - D(X) - D(Y) \quad (3)$$

$$\left(M = \sum_{i=1}^{t} m_i, N = \sum_{i=1}^{t} n_i\right)$$

where $ID(X, Y)$ is a function of two sources, and $D(X + Y)$ is the measure of diversity of the mixed source $D(X + Y) = (M + N)\log(M + N) - M\log M - N\log N$. It can be proved that the increment of diversity satisfies $0 \le ID(X, Y) \le D(X + Y)$.

Given a problem of classification of sequence X, we shall compare X with a standard set (called standard source), which consists of samples with known properties. The standard source of diversity S is defined by

$$D(S) = D(m_1, m_2, \cdots, m_t) = M\log M - \sum_{i=1}^{t} m_i \log m_i \quad (4)$$

$$\left(M = \sum_{i} m_i\right)$$

where m_i is the sum of the frequency of the i-th category of the feature variables over all samples in the standard set. The increment of diversity [$ID(X, S)$] between X and S is still given by Eq. 3. The increment of diversity of two sources is essentially a measure of their similarity level. The smaller $ID(X, S)$, the higher the similarity level between X and S.

Quadratic discriminant analysis (QD)

For a sequence X to be classified, the increment of diversity between source X and the standard source is denoted by an r-dimensional vector, called R. r increments of diversity are integrated by using the quadratic discriminant function (Zhang 1997; Zhang and Luo 2003). Here we shall generalize the quadratic discriminant formulation for the classification of multi-groups. Consider X classified into n groups (ω_1,ω_2,...ω_n). The discriminant function between group i and group j is defined by

$$\xi_{ij} = \ln p(\omega_i|x) - \ln p(\omega_j|x). \quad (5)$$

According to Bayes' theorem, we can deduce the following equation for the two-group case (see Appendix 2)

$$\xi_{ij} = \ln\frac{p_i}{p_j} - \frac{\delta_i - \delta_j}{2} - \frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}. \quad (6)$$

The result can be generalized to n-groups directly. Set

$$\eta_v = \ln p_v - \frac{\delta_v}{2} - \frac{1}{2}\ln|\Sigma_v| \quad (7)$$

$$\delta_v = (R - \mu_v)\ \Sigma_v^{-1}\ (R - \mu_v) \quad (v = 1, \ldots, n) \quad (8)$$

where p_v denotes the number of samples in group v, μ_v denotes increments of diversity averaged over group v, $|\Sigma_v|$ is the determinant of matrix Σ_v and δ_v is the square Mahalanobis distance between R and μ_v with respect to Σ_v (Note that μ_v and $|\Sigma_v|$ are calculated in training set).

From Eqs. 6 and 7, we have

$$\xi_{ij} = \eta_i - \eta_j \quad (i,j = 1, \cdots, n). \quad (9)$$

It can be easily proved that $p(\omega_k|X)$ is the maximum of $p(\omega_v|X)$ if η_k is the maximum in η_v (v = 1,...,n). So, we predict that X belongs to group k.

Correction to IDQD prediction

The section includes two steps: (1) The structure fluctuation is removed. It is called the structure fluctuation in prediction if the predicted structure of one residue is different from its left and right neighbors, and the two neighbors are predicted as the same structure. The structure of central residue for the tri-peptide should be corrected to the same as the prediction of its left and right neighbors. (2) The structure boundary is corrected by using tetra-peptide boundary words after removal of fluctuation. There are four kinds of boundary words, namely, α\c-type boundary words (including subtypes "ααα c," "αα cc," and "α ccc"); β\c-type boundary words ("βββ c," "ββ cc," and "β ccc"); c\α-type boundary words ("ccc α," "cc αα," and "c ααα"); and c\β-type boundary words ("ccc β," "cc ββ," and "c βββ"). If the tetra-peptide on a predicted structural boundary, for example, α\c boundary, is in full accordance with three boundary words of "ααα c," "αα cc," and "α ccc" subtype, then the score = 3 is made in the prediction; if it is partly in accordance with three boundary words, namely, in accordance with only two or one boundary words of three subtypes then the score = 2 or 1; if no accordance at all, then the score = 0. Next, we assume boundary shifting toward left or right by one or two residues. In each case we score the prediction. Finally, we choose the maximum score as the ultimately predicted structural boundary through comparison of the above five cases. If the same score is obtained for several cases, we always choose the one with the least shift as the predicted.

Results

Tetra-peptide structural words at different confidence levels in CB513 dataset

Three kinds of tetra-peptide structural words and 12 kinds of tetra-peptide boundary words with different confidence levels have been deduced by using Eq. 1. Generally, the structural words with high confidence level give more accurate results. However, the number of these words is too small to afford enough information. In contrast, the serious overlap between different types of structure words with low confidence levels leads to ambiguity in prediction. Therefore, using appropriate structural words would yield a prediction with higher accuracy. We have utilized alpha and beta words at 80% confidence level and coil words at 55% confidence level. There is less overlap among the three types of structural words. To ensure a high enough number of structural words, we utilized coil words with the lower confidence level since coil itself is an irregular structure. Similarly we selected boundary words at 55% confidence level in the correction program of the algorithm. The numbers of tetra-peptide structural words with different confidence levels are summarized in Table 1, and the details of tetra-peptide structural words used in our TPIDQD algorithm can be found in the "Electronic Supplementary Material 1–7".

The secondary-structure prediction of the central residue for a 21-residue fragment

We have predicted the structure of the central residue of a 21-residue fragment in the middle of a protein sequence by using the TPIDQD method (namely, central residues are located in the sequence from the 11th site at N′ terminal to the 11th site at C′ terminal). A set of 21-residue fragments was obtained by sliding a window of 21 residues along all protein sequences. They compose set A, B, or C in terms of the structure classes of their central residues. Each 21-residue fragment is equally divided into three segments, denoted as left (L), middle (M), and right (R) segments. By calculating the frequency of alpha words, beta words, and coil words occurring in L, M, and R segments of all samples in set A and using Eq. 4, we obtain nine standard sources of diversity (nine-dimensional vector). Similarly, nine standard sources of diversity on set B and set C can be deduced. For the 21-residue fragment X to be predicted, we can obtain nine increments of diversity between source X and $S(S = A, B, C)$ by using Eq. 3. Then we calculate δ_v and $|\Sigma_v|(v = A, B, C)$ in the training set. Following the quadratic discriminant function and using Eq. 7, we find the maximum among η_A, η_B and η_C; then we predict that X belongs to its structure class. Similarly, we have predicted the structures of first (last) 10 residues at the N′ (C′) end of the chain by using 21-residue fragments. (Note that the addition of several blanks at the terminal is required to obtain the 21-residue fragment.) Since the IDQD is a method of statistical prediction, the fluctuation exists inevitably. In particular, those residues in the boundary of a secondary structure may easily be misidentified. Therefore, we have introduced the correction program to the IDQD prediction as given in the "Materials and methods" section.

Among the independent dataset tests, the k-fold cross-validated test and the jackknife test have often been used for examining the accuracy of a prediction method. The jackknife test is deemed the most rigorous and objective (Chou and Shen 2007b, 2008; Chou and Zhang 1995) and is widely and increasingly used to test the power of various statistical prediction methods (Cao et al. 2006; Chen et al. 2006a, b, 2007; Chen and Li 2007; Diao et al. 2007, 2008; Du and Li 2006; Fang et al. 2008; Gao and Wang 2006; Gao et al. 2005; Guo et al. 2006a, b; Huang and Li 2004; Jahandideh et al. 2007; Kedarisetti et al. 2006; Li and Li 2008; Lin and Li 2007a, b; Mondal et al. 2006; Mundra et al. 2007; Pugalenthi et al. 2007; Shi et al. 2007; Sun and Huang 2006; Tan et al. 2007; Wen et al. 2006; Zhang et al. 2006a; Zhang and Ding 2007; Zhang et al. 2006b; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007). However, considering the longer time needed for the jackknife test and because the goal of our paper concentrated on introducing a new method for structure prediction and comparing it with published results, we adopt the threefold cross-validation to evaluate the prediction quality. We divided the training set randomly into three parts, two of which were for training and the rest for testing. The process was repeated three times.

Table 1 The number of tetra-peptide words of 15 structural types

Structural type	Confidence level						
	99%	95%	85%	80%	75%	65%	55%
$\alpha\alpha\alpha\alpha$	118	430	1,747	2,385	2,394	2,660	6,076
$\beta\beta\beta\beta$	72	534	653	2,378	2,379	3,066	3,461
$cccc$	38	280	1,372	1,885	1,890	2,124	5,912
$\alpha\alpha cc$	89	117	1,214	1,339	1,346	1,347	1,347
$\beta\beta cc$	89	97	1,354	1,431	1,439	1,439	1,439
$cc\alpha\alpha$	95	99	1,218	1,242	1,244	1,244	1,244
$cc\beta\beta$	124	139	1,508	1,607	1,612	1,614	1,614
$\alpha\alpha\alpha c$	92	126	1,430	1,562	1,579	1,579	1,579
αccc	67	74	1,141	1,177	1,177	1,177	1,177
$\beta\beta\beta c$	80	93	1,251	1,320	1,325	1,325	1,325
βccc	56	58	1,132	1,158	1,161	1,161	1,161
$ccc\alpha$	61	61	1,052	1,059	1,061	1,061	1,061
$c\alpha\alpha\alpha$	128	134	1,344	1,423	1,427	1,427	1,427
$ccc\beta$	73	82	1,268	1,294	1,296	1,296	1,296
$c\beta\beta\beta$	96	108	1,295	1,373	1,379	1,379	1,379

The first three rows give the number of alpha, beta, and coil words. The next rows present boundary words

Table 2 Prediction accuracy for protein secondary structure

21-Residue fragment position (test)	Q_3 (%)	Q_3^* (%)	Q_3^{corr} (%)
Fragment in middle (direct)	71.64	76.89	82.78
Fragment in middle (threefold cross-validation)	63.18	70.90	78.67
Fragment at 5′ end (direct)	73.75	78.97	84.12
Fragment at 5′ end (threefold cross-validation)	69.87	74.02	81.19
Fragment at 3′ end (direct)	72.16	78.05	83.46
Fragment at 3′ end (threefold cross-validation)	67.92	73.27	80.01
Fragment in full sequence (direct)	72.30	77.13	83.07
Fragment in full sequence (threefold cross-validation)	66.82	72.38	79.19

Q_3 denotes the Q_3 score of IDQD prediction without correction, Q_3^* denotes the Q_3 score after fluctuation is removed, Q_3^{corr} denotes the Q_3 score after fluctuation removal and boundary correction. Two kinds of tests are given, "direct" means direct application of TPIDQD trained by 513 proteins

All results (IDQD predictions and corrections in the middle of all chains, at the N′ and C′ terminals, and in full protein chains) are listed in Table 2.

The secondary-structure prediction of central residues for different fragments

In the TPIDQD algorithm, we constructed sources of diversity and calculated the measure of diversity. One increment of diversity corresponded to one feature variable. Nine feature variables were assumed, and each variable was defined in a segment of seven residues (corresponding to a 21-residue fragment) in the above calculation. By using the same approach, we examined the effect of different segment lengths, namely the segment of five residues (corresponding to 15-residue fragment), nine residues (27-residue fragment) and 11 residues (33-residue fragment), on the accuracy of secondary-structure prediction and found that the best prediction is given in the case of the seven-residue segment. The results are listed in Table 3.

Table 3 The secondary-structure prediction for different length segments

N_R	N_S	Q_3 (%)
15 residues	3	63.38
21 residues	3	71.64
27 residues	3	71.38
33 residues	3	68.97

N_R Number of residues in a fragment for structure prediction, N_S number of segments into which each fragment is divided. The predictions are given by direct application of IDQD trained by 513 proteins (compared with Table 2)

The influence of long-range amino acid interactions on the secondary-structure formation of proteins is a complex problem (Kihara 2005; Tsai and Nussinov 2005). However, in our method we can easily take long-range sequence information into account by increasing the fragment length. The fragment length of 21 residues was assumed in previous sections, and the long-range information has been partially considered there. To further consider the long-range sequence information, as an example, we test the 49-residue fragment (seven segments of seven residues instead of the three segments above) and define 21 increments of diversity as the feature variables (instead of nine variables above). We find that the Q_3 scores are several points higher than those given in Table 2. Therefore, the TPIDQD algorithm has the capability of calculating the influence of long-range residue interactions and largely improving the secondary-structure prediction through introduction of longer fragments.

Discussion

In the TPIDQD algorithm, a high enough number of tetra-peptide structural words is a key factor for the success of the algorithm. In defining structural words, we only consider tetra-peptides with $N \geq 2$ in the dataset, since the occurrence once of a tetra-peptide in a given secondary structure may be at random. However, the probability of its occurrence two or more times at random in the same secondary structure is very small. The probability of random occurrence twice is about 1/81 and three times is $(1/81)^2$. Therefore the tetra-peptide word with higher confidence level under this constraint ($N \geq 2$) should really be a code word characteristic of a certain structure. However, we find a lot of tetra-peptides occurring only once in the CB513 dataset. If the condition $N \geq 2$ is not required, then one would obtain more tetra-peptide structural words following Eq. 1. When we use three kinds of tetra-peptide words (under condition $N \geq 1$) as the diversity sources to predict the structure of the same 21-residue fragment, the Q_3 score is again higher than that of the prediction based on $N \geq 2$ tetra-peptide words. (The values of $Q_3 = 82.75$ and 78.30% were achieved in direct test and threefold cross-validation test respectively under condition $N \geq 1$). The reason is that one can obtain more structural words at higher confidence levels under condition $N \geq 1$, and this makes the prediction more accurate.

Although tetra-peptide structural words occurring only once in a dataset may be an accident, and we have ignored them in the model of $N \geq 2$, these results show that a large number of structural words at higher confidence level are an important factor for improving the performance of prediction. Note that the total number of tetra-peptides is 160,000, and a portion of them are candidates for code words useful in structure determination. So there exists an

upper limit of the number of tetra-peptide structural words. We have utilized 25,513 tetra-peptide structural words ($N \geq 2$) based on the CB513 dataset in this paper. With the rapid expansion of the protein structure dataset, more tetra-peptide words with a higher confidence level will be obtainable, making the prediction more accurate.

The prediction of protein secondary structure is one of the typical classification problems in bioinformatics. We obtained an overall Q_3 score of 79.19% for secondary-structure prediction using the TPIDQD algorithm. The accuracy can be further improved by taking long-range sequence information (fragments with more than 21 residues) into account in prediction. The prediction capability of the TPIDQD method is comparable with that of other currently published top algorithms. In the same CB513 dataset, for example, the dual-layer SVM algorithm attained an overall Q_3 score of 75.2% (Guo et al. 2004), the SPINE yielded a tenfold cross-validated accuracy of $Q_3 = 76.77\%$ (Dor and Zhou 2007), and the dynamic programming algorithm achieved an overall Q_3 score of 66% (Sadeghi et al. 2005). The results show that tetra-peptide signals can indeed reflect some relationship between an amino acid's sequence and its secondary structure, indicating the importance of tetra-peptide structural words in protein structure prediction.

Acknowledgments The work was supported by the National Science Foundation of China (No. 90403010). The authors are grateful to Drs. Jun Lu, Ying Zhang, and Hao Lin for their helpful discussions.

Appendix 1

Table 4

Table 4 Probability parameters in defining tetra-peptide structural words in CB513 dataset

j	m_j	p_j
$\alpha\alpha\alpha\alpha$	19,741	0.24880
$\beta\beta\beta\beta$	8,247	0.10394
$cccc$	19,601	0.24704
$\alpha\alpha cc$	2,627	0.03311
$\beta\beta cc$	2,909	0.03666
$cc\alpha\alpha$	2,481	0.03127
$cc\beta\beta$	3,069	0.03868
$\alpha\alpha\alpha c$	3,025	0.03812
αccc	2,226	0.02806
$\beta\beta\beta c$	2,751	0.03467
βccc	2,327	0.02933
$ccc\alpha$	2,070	0.02609
$c\alpha\alpha\alpha$	2,945	0.03712
$ccc\beta$	2,514	0.03168
$c\beta\beta\beta$	2,811	0.03543

$M = \sum_j m_j$=79,344; $p_j = \frac{m_j}{M}$; $\sum_j p_j$=1

Appendix 2: the deduction of quadratic discriminant analysis in two-group case by using Bayesian theorem

For a sequence X to be classified between group ω_1 and group ω_1, assuming ω_1 is positive set and ω_2 is negative set, the discriminant function is defined by

$$\xi = \ln p(\omega_1|x) - \ln p(\omega_2|x). \tag{10}$$

According to Bayes' theorem,

$$p(\omega_l|x) = p(\omega_l)p(x|\omega_l)/p(x) \quad (l = 1, 2) \tag{11}$$

where $p(\omega_l)$ is the probability a priori of set l (l = 1, 2), inserting Eq. 11 into Eq. 10, we obtain

$$\xi = \ln\frac{p(\omega_1)}{p(\omega_2)} + \ln\frac{p(x|\omega_1)}{p(x|\omega_2)}. \tag{12}$$

Assume normal distribution of feature variables (M-dimensional vector) in two sets

$$p(x|\omega_l) = \frac{1}{Z_l}\exp\left(-\frac{1}{2}(x-\mu_l)^T\sum_l^{-1}(x-\mu_l)\right) \tag{13}$$

$$Z_l = (2\pi)^{M/2}|\Sigma_l|^{1/2} \quad (l = 1, 2)$$

where μ_l (M-dimensional vector) and Σ_l ($M \times M$ matrix) are the mean and covariant of feature variables over positive and negative sets respectively. Inserting Eq. 13 into Eq. 12, we obtain

$$\begin{aligned}\xi = \ln\frac{p(\omega_1)}{p(\omega_2)} - \frac{1}{2}((x-\mu_1)^T\sum_1^{-1}(x-\mu_1) \\ - (x-\mu_2)^T\sum_2^{-1}(x-\mu_2)) - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|}\end{aligned} \tag{14}$$

$$\text{Set} \quad \delta_l = (x-\mu_l)\sum_l^{-1}(x-\mu_l) \quad (l = 1,\ 2) \tag{15}$$

$$\text{So} \quad \xi_{ij} = \ln\frac{p_i}{p_j} - \frac{\delta_i - \delta_j}{2} - \frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}. \tag{16}$$

This result can easily be generalized to more than two groups as shown in text.

References

Cai YD, Li YX, Chou KC (2000) Using neural networks for prediction of domain structural classes. Biochim Biophys Acta 1476:1–2

Cao Y, Liu S, Zhang L et al (2006) Prediction of protein structural class with rough sets. BMC Bioinformatics 7:20

Cedano J, Aloy P, P'erez-Pons JA et al (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600

Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381

Chen C, Tian YX, Zou XY et al (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448
Chen C, Zhou X, Tian Y et al (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357:116–121
Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428
Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1) D amino acid composition space. Proteins 21:319–344
Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255 (Erratum: ibid., 2001, 44:60)
Chou KC (2005a) Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6(5):423–436
Chou KC (2005b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou KC, Cai YD (2006) Predicting protein-protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322
Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12:107–118
Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13:211–215
Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11:523–538
Chou KC, Shen HB (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345
Chou KC, Shen HB (2007b) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Shen HB (2007c) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640
Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162
Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–22020
Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34:508–519
Diao Y, Li M, Feng Z et al (2007) The community structure of human cellular signaling network. J Theor Biol 247:608–615
Diao Y, Ma D, Wen Z et al (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids 34:111–117
Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815
Dor O, Zhou Y (2007) Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 66:838–845
Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7:518
Fang Y, Guo Y, Feng Y et al (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34:103–109
Frishman D, Argos P (1995) Knowledge-based secondary structure assignment. Proteins 23:566–579
Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19:511–516
Gao QB, Wang ZZ, Yan C et al (2005) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579:3444–3448
Garnier J, Osguthorpe DJ, Robson B (1978) Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120
Guo J, Hu C, Sun ZR et al (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins 54:738–743
Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6:5099–5105
Guo YZ, Li M, Lu M et al (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402
Hua SJ, Sun ZR (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308:397–407
Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20:21–28
Jahandideh S, Abdolmaleki P, Jahandideh M et al (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128:87–93
Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637
Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856
Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988
Kihara D (2005) The effect of long-range interactions on the secondary structure formation of proteins. Protein Sci 14:1955–1963
Laxton RR (1978) The measure of diversity. J Theor Biol 71:51–67
Li FM, Li QZ (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125
Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213:493–502
Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354:548–551
Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466
Lin K, Simossis VA, Taylor WR et al (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21:152–159
Liu H, Wang M, Chou KC (2005) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739
Mondal S, Bhavna R, Mohan Babu R et al (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243:252–260

Montgomerie S, Sundararaj S, Gallin WJ et al (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. Bioinformatics 7:301–313

Mundra P, Kumar M, Kumar KK et al (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recognit Lett 28:1610–1615

Pauling L, Corey RB, Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37:205–234

Petersen TN, Lundegaard C, Nielsen M et al (2000) Prediction of protein secondary structure at 80% accuracy. Proteins 41:17–20

Pugalenthi G, Tang K, Suganthan PN et al (2007) A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. BMC Bioinformatics 8:351

Rackovsky S (1993) On the nature of protein folding code. Proc Natl Acad Sci USA 90:644–648

Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 3:71–84

Rost B, Sander C (1993) Prediction of secondary structure at better than 70% accuracy. J Mol Biol 232:584–599

Rost B, Sander C (2000) Third generation prediction of secondary structures. Methods Mol Biol 143:71–95

Sadeghi M, Parto S, Arab S et al (2005) Prediction of protein secondary structure based on residue pair types and conformational states using dynamic programming algorithm. FEBS Lett 579:3397–3400

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

Shen HB, Chou KC (2007b) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363:297–303

Shi JY, Zhang SW, Pan Q et al (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475

Tan F, Feng X, Fang Z et al (2007) Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids 33:669–675

Tsai CJ, Nussinov R (2005) The implications of higher (or lower) success in secondary structure prediction of chain fragments. Protein Sci 14:1943–1944

Wang M, Yang J, Liu GP et al (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Sel 17:509–516

Wang M, Yang J, Xu ZJ et al (2005) SLLE for predicting membrane protein types. J Theor Biol 232:7–15

Ward JJ, McGuffin LJ, Jones DT (2003) Secondary structure prediction with support vector machines. Bioinformatics 19:1650–1655

Wen Z, Li M, Li Y et al (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32:277–283

Xiao X, Shao SH, Ding YS et al (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30:49–54

Xu KX (1999) Biomathematics (in Chinese). Science Press, Beijing

Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA 94:565–568

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang LR, Luo LF (2003) Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res 31:6214–6220

Zhang SW, Pan Q, Zhang HC et al (2006a) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30:461–468

Zhang ZH, Wang ZH, Zhang ZR et al (2006b) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174

Zhang TL, Ding YS, Chou KC (2007) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44:57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50:44–48

Zhou XB, Chen C, Li ZC et al (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551